第一天day1：hadoop环境搭建

准备：

0、安装MobaXterm\_Setup\_9.2：用于windows以ssh方式连接linux系统

MobaXterm可以实现文档的上传下载，就不需要设置虚拟机与宿主机之间的共享

1、安装虚拟机VMware\_workstation\_full\_12.5.2：用于安装多个linux系统

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*已经废弃，原因是centos门槛太高，学习进度缓慢\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

2、安装centos7 CentOS-7-x86\_64-DVD-1511

centos01

ip：192.168.255.128

为hadoop用户增加管理员权限时出错：

问题 1、下面语句不管用

$ sudo adduser hadoop sudo #为hadoop用户增加管理员权限

解决方案：手动修改visudo

1、root用户下执行命令：visudo

2、在root ALL=(ALL) NOPASSWD:ALL

下面添加一行：hadoop ALL=(ALL) NOPASSWD:ALL

保存即可

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

2、安装ubuntu ubuntu-16.04-desktop-amd64

虚拟机安装ubuntu过程简单，自己找网上教程

注意两点

1、安装的时候不要联网，否则要下载许多其他软件，太慢了

2、用户名/密码 tsing01/123

安装完成后，进行以下操作：

1. 选择网络适配器模式：NAT

VMnet1是使用的是host-only模式，即虚拟机只能与主机构成内部通信，无法对外网进行通信。VMnet8使用的是NAT网络模式，这种模式一般是一台宿主机以及多台虚拟机，在构建hadoop分布式集群的时候就采用这种模式即可

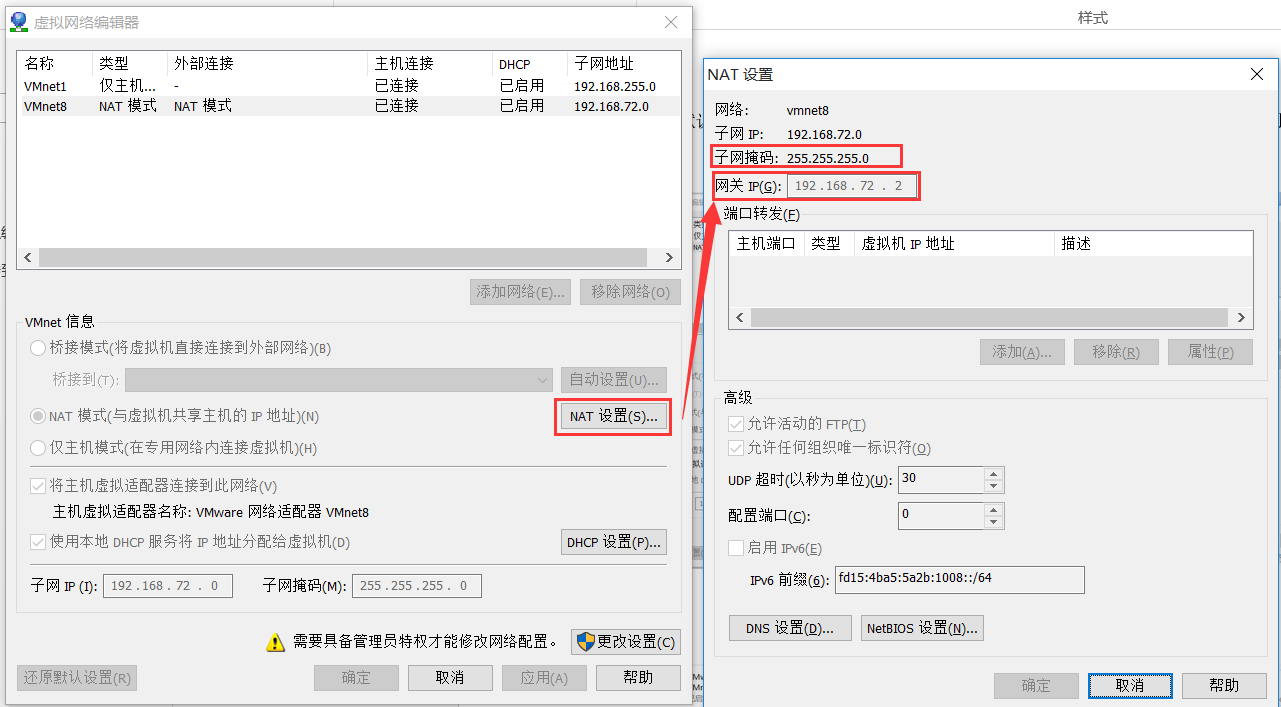
1. 更新源，安装vim

$ sudo apt upgrade //升级

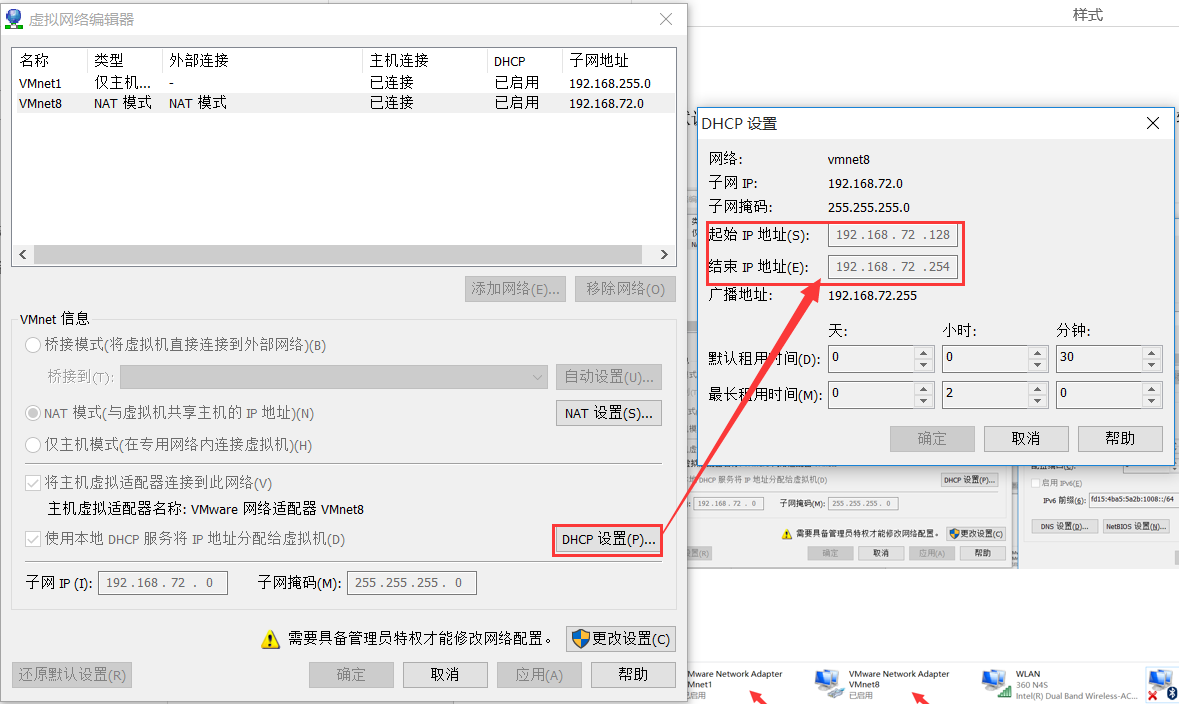
$ sudo apt update //更新 这两个顺序不可颠倒

$ sudo apt install vim //安装vim

1. 设置静态IP：固定的ip便于ssh连接和hadoop的相关设置
2. 获取默认网关：打开VM：选中要查看网卡的虚拟机 -> 编辑 -> 虚拟网络编辑器



获取可以设置的ip的范围



1. 设置ip：vim 编辑interface

$ sudo vim /etc/network/interfaces

追加几行代码：

# the primary network interface

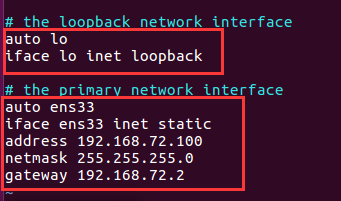
auto ens33 //默认网卡

iface ens33 inet static //静态网络

address 192.168.72.130 //ip地址

netmask 255.255.255.0 //子网掩码

gateway 192.168.72.2 //网关



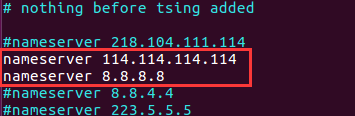
1. 设置DNS：vim编辑

$ sudo vi /etc/resolvconf/resolv.conf.d/base

追加几行代码：

nameserver 8.8.8.8

nameserver 8.8.4.4



常用DNS有：

在宿主机使用ping命令验证可用的DNS

googleDNS： 8.8.8.8 8.8.4.4

114DNS： 114.114.114.114 114.114.115.115

阿里DNS： 223.5.5.5 223.6.6.6

中科大的DNS

202.38.64.1

202.112.20.131

202.141.160.95

202.141.160.99

202.141.176.95

202.141.176.99

1. 重启网络服务

sudo /etc/init.d/networking restart #使网卡配置生效

sudo /etc/init.d/resolvconf restart #使DNS生效

如果无效就重启系统

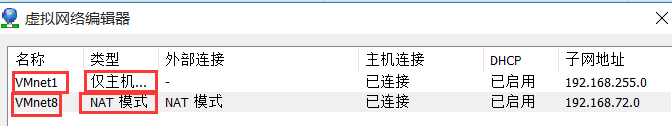
\*注意 自从ubuntu15.04网络管理工具废弃了nm-tool，改为nmcli

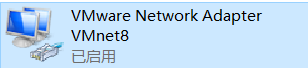
nmcli dev show //查看mac ip DNS

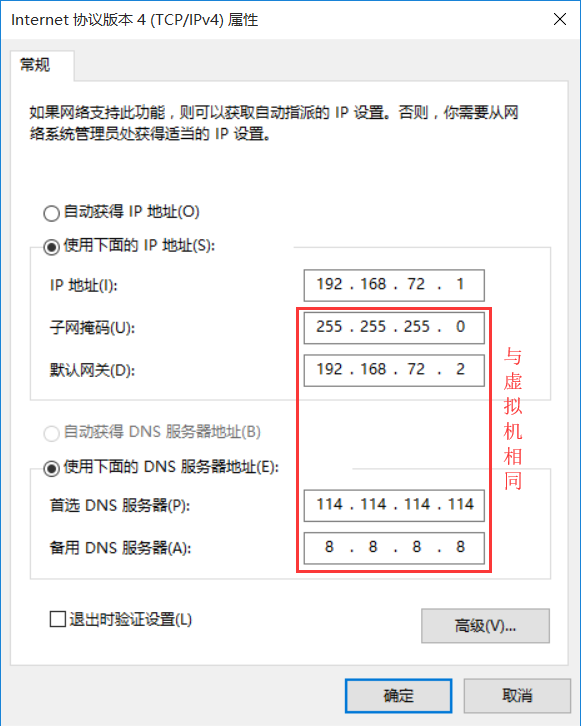
比较优秀的教程 <https://www.cnblogs.com/EasonJim/p/7857671.html>

1. 设置宿主机虚拟网卡

找到该网络类型对应的虚拟网卡：



例如：NAT，需改的ip为静态的：



ubuntu搭建hadoop教程地址：https://www.cnblogs.com/87hbteo/p/7606012.html

3、创建hadoop用户：

$ sudo useradd -m hadoop -s /bin/bash #创建hadoop用户，并使用/bin/bash作为shell

$ sudo passwd hadoop #为hadoop用户设置密码，之后需要连续输入两次密码

$ sudo adduser hadoop sudo #为hadoop用户增加管理员权限

$ su - hadoop #切换当前用户为用户hadoop

$ sudo apt-get update #更新hadoop用户的apt,方便后面的安装

4、安装java

默认安装了openjdk，因为openjdk是jdk的精简版本，阉割了许多功能，还是需要安装jdk

查看是否安装openjdk

$ java -version

卸载openjdk

$ sudo apt autoremove openjdk\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*放弃 原因：速度慢 时间长\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

添加ppa源

$ sudo add-apt-repository ppa:webupd8team/java

更新源数据库

$ sudo apt-get update

安装 Oracle Java 8

$ sudo apt-get install oracle-java8-installer

查看java版本

$ java -version

设置java环境变量

$ sudo apt-get install oracle-java8-set-default

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

下载jdk-8u111-linux-x64.tar 放到Downlaods目录下

$ cd ~/Downlaods

解压tar包

$ sudo tar -zxvf jdk-8u111-linux-x64.tar.gz -C /usr/lib

进入目录

$ cd /usr/lib

重命名为java，如果jvm不存在， sudo mkdir jvm

$ sudo mv jdk1.8.0\_111 jvm/java

备份配置文件，防止改动失误

$ cp ~/.bashrc ~/.bashrc.copy

vim编辑配置文件

$ vim ~/.bashrc

在.bashrc文档尾部追加四行代码

export JAVA\_HOME=/usr/lib/jvm/java

export JRE\_HOME=${JAVA\_HOME}/jre

export CLASSPATH=.:${JAVA\_HOME}/lib:${JRE\_HOME}/lib

export PATH=${JAVA\_HOME}/bin:$PATH

如果懒得手动输入代码，可以选择

1）重定向

将代码写入一个文件 a

$ cat a >> ~/.bashrc

2）安装vim-gnome

$ sudo apt install vim-gnome

在vim界面输入 "+p 完成从系统粘贴板到文档的复制

附加：

"+y ：复制到粘贴板

使配置文件立即生效

$ source ~/.bashrc

验证安装成功？

$ java -version

5、安装hadoop

先下载hadoop-2.8.0.tar.gz，链接如下:

http://mirrors.hust.edu.cn/apache/hadoop/common/

下面进行安装：

$ sudo tar -zxvf hadoop-2.8.0.tar.gz -C /usr/local #解压到/usr/local目录下

$ cd /usr/local

$ sudo mv hadoop-2.8.0 hadoop #重命名为hadoop

$ sudo chown -R hadoop ./hadoop #修改文件权限

给hadoop配置环境变量，将下面代码添加到.bashrc文件:

export HADOOP\_HOME=/usr/local/hadoop

export CLASSPATH=$($HADOOP\_HOME/bin/hadoop classpath):$CLASSPATH

export HADOOP\_COMMON\_LIB\_NATIVE\_DIR=$HADOOP\_HOME/lib/native

export PATH=$PATH:$HADOOP\_HOME/bin:$HADOOP\_HOME/sbin

6、伪分布式配置

Hadoop 可以在单节点上以伪分布式的方式运行，Hadoop 进程以分离的 Java 进程来运行，节点既作为 NameNode 也作为 DataNode，同时，读取的是 HDFS 中的文件。Hadoop 的配置文件位于 /usr/local/hadoop/etc/hadoop/ 中，伪分布式需要修改2个配置文件 core-site.xml 和 hdfs-site.xml 。Hadoop的配置文件是 xml 格式，每个配置以声明 property 的 name 和 value 的方式来实现。首先将jdk1.7的路径添（export JAVA\_HOME=/usr/lib/jvm/java ）加到hadoop-env.sh文件

接下来修改core-site.xml文件：

<configuration>

<property>

<name>hadoop.tmp.dir</name>

<value>file:/usr/local/hadoop/tmp</value>

<description>Abase for other temporary directories.</description>

</property>

<property>

<name>fs.defaultFS</name>

<value>hdfs://localhost:9000</value>

</property>

</configuration>

接下来修改配置文件 hdfs-site.xml

<configuration>

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

<property>

<name>dfs.namenode.name.dir</name>

<value>file:/usr/local/hadoop/tmp/dfs/name</value>

</property>

<property>

<name>dfs.datanode.data.dir</name>

<value>file:/usr/local/hadoop/tmp/dfs/data</value>

</property>

</configuration>

Hadoop 的运行方式是由配置文件决定的（运行 Hadoop 时会读取配置文件），因此如果需要从伪分布式模式切换回非分布式模式，需要删除 core-site.xml 中的配置项。此外，伪分布式虽然只需要配置 fs.defaultFS 和 dfs.replication 就可以运行（可参考官方教程），不过若没有配置 hadoop.tmp.dir 参数，则默认使用的临时目录为 /tmp/hadoo-hadoop，而这个目录在重启时有可能被系统清理掉，导致必须重新执行 format 才行。所以我们进行了设置，同时也指定 dfs.namenode.name.dir 和 dfs.datanode.data.dir，否则在接下来的步骤中可能会出错。

配置完成后，执行 NameNode 的格式化

$ ./bin/hdfs namenode -format

启动namenode和datanode进程，并查看启动结果

$ ./sbin/start-dfs.sh

$ jps

启动完成后，可以通过命令 jps 来判断是否成功启动，若成功启动则会列出如下进程: “NameNode”、”DataNode” 和 “SecondaryNameNode”

7、克隆虚拟机

将配置好的虚拟机文件拷贝2份，然后虚拟机中文件 -> 打开 选择拷贝的文件；在打开备份的虚拟机时会弹出对话框，选择复制就好。

ubuntu00作为主机00

ubuntu01作为副机01

ubuntu02作为副机02

8、Linux多台机器配置ssh免登录

1.安装ssh. sudo apt-get install ssh. 安装完成后会在~目录（当前用户主目录，即这里的/home/xuhui）下产生一个隐藏文件夹.ssh（ls -a 可以查看隐藏文件）。如果没有这个文件，自己新建即可（mkdir .ssh）.

2.进入.ssh目录下面，在每台机器上执行：ssh-keygen -t rsa 之后一路回车，产生密钥；

3.完成第二步后会产生两个文件：

id-rsa #私钥

id-rsa.pub #公钥

4.在第一台机器的目录.ssh下执行命令，$ cat id-rsa.pub >> authorized\_keys；此后.ssh下面会出现authorized\_keys文件。

5.然后将第一台机器的.ssh目录下面的authorized\_keys文件拷贝到第二台计算机的.ssh目录下，如：scp authorized\_keys xuhui@192.168.72.101:~/.ssh/

6.再转到第二台机器的.ssh目录下，会发现刚刚传输过来的文件-authorized\_keys，然后执行命令，将第二台计算机的公钥也加进来，如：cat id-rsa.pub >> authorized\_keys.

7.将第二台计算机新生成的authorized\_keys传输第三台计算机，将第三台计算机的公钥-id-rsa.pub添加到从第二台计算机传过来的authorized\_keys里面。

8.依次类推，直至最后一台计算机。

9.在最后一台计算机执行完添加后，生成的authorized\_keys文件就包含所有计算机的公钥，如果以后还有机器加进来，可以直接添加到文件-authorized\_keys。最后，将最后生成的authorized\_keys复制到每一台计算机的.ssh目录下，覆盖掉之前的authorized\_keys。

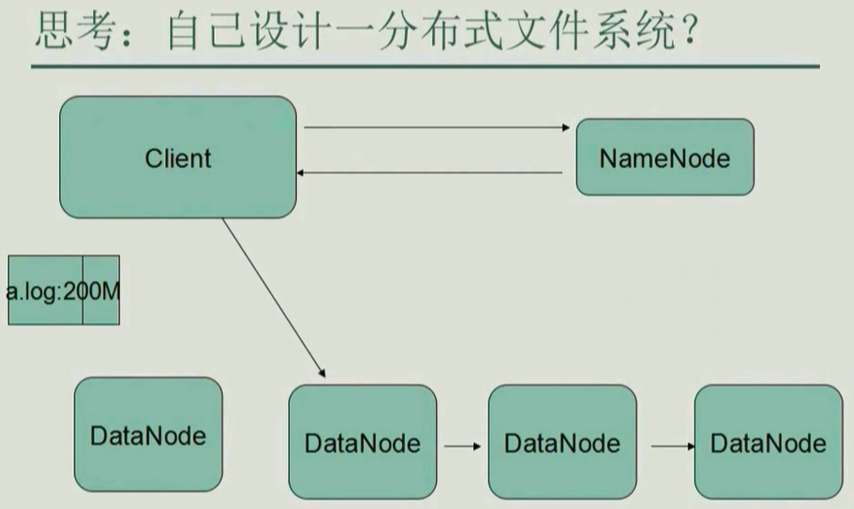
10.完沉第九步后，就可以在任意一台计算机上，免密码ssh登录到其他计算了。

上面十步是笨法，简单方法是使用ssh-copy-id：

hadoop@ubuntu00:~$ ssh-copy-id root@ubuntu02 # 用户hadoop@ubuntu00的公钥赋给用户root@ubuntu02。Ssh是单向的，现在可以实现root@ubuntu02免密登陆hadoop@ubuntu00,但不能反过来免密，还需

root@ubuntu02:~# ssh-copy-id hadoop@ubuntu00

# 第二天 day2：



Hadoop2 一块128M 为什么分块？

写一块 成功一块

读一块 成功一块

读写容错性高

Pip line流水线

分布文件系统：网络共享、透明性、容错性

常见的分布式文件系统：GFS HDFS mogoDB Ceph FastDF TFS

HDFS：一次写入 多次查询，不支持并发（一个文件的多个块不可以一起写入）；小文件不合适（1M的文件也占用一块128M）

1. HDFS的shell：通过命令行操作hdfs系统的文件

格式化配置HDFS文件系统，打开NameNode(HDFS服务器)，然后执行以下命令。

$ hadoop namenode -format

进入hadoop工作目录

$ cd /usr/local/Hadoop

启动HDFS

$ ./sbin/start-dfs.sh

$ jps

$ Hadoop fs –ls –R –h / # ls:列表 –R：递归列表 –H：人类可看（显示文件大小）

$ Hadoop fs -copyFromLocal

$ Hadoop fs -copyToLocal #

$ Hadoop fs –get # 下载到ubuntu本地

$ Hadoop fs –rm -R #删除

$ Hadoop fs –count #统计

$ hadoop fs -mkdir /tsing01 #创建目录

$ hadoop fs –chmod +x /tsing01 #加执行权限

$ hadoop fs –chmod -x /tsing01 #去执行权限

$ hadoop fs –chown root /tsing01 #改变所属用户

$ hadoop fs –chgrp root /tsing01 #改变所属组

$ hadoop fs –chown root:supergroup /tsing01 #改变所属用户:组

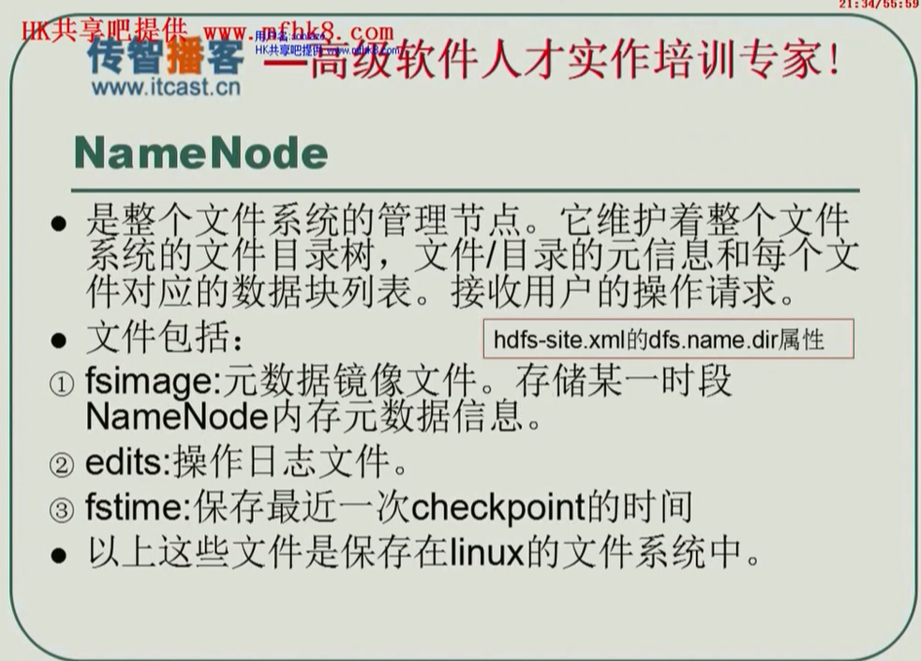
$ hdfs dfs可以替换 Hadoop fs

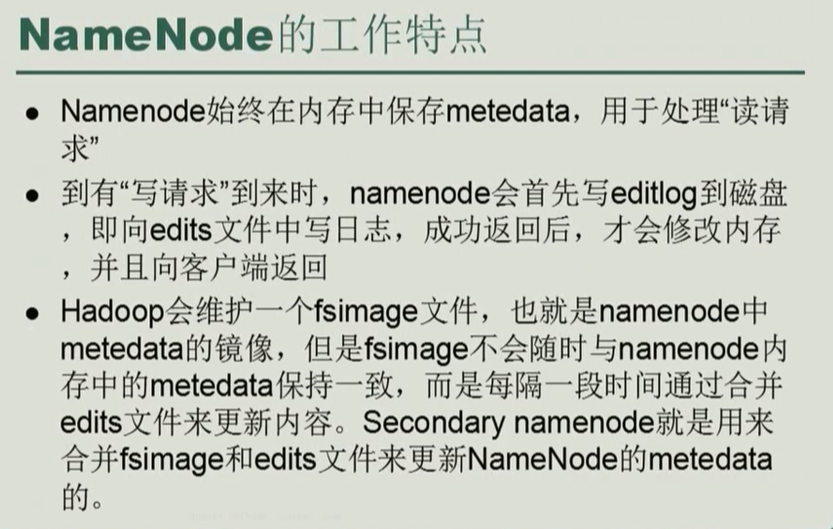
1. HDFS架构：

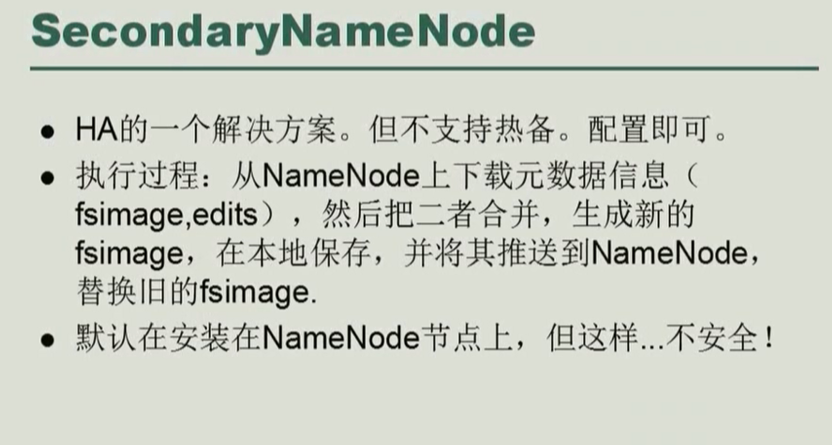
元数据的存储细节 Name metadata



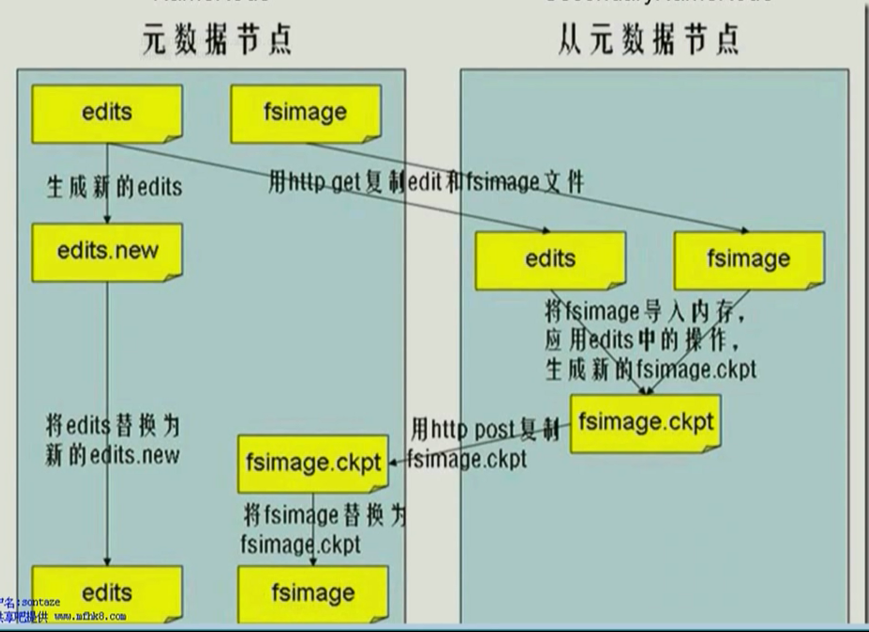
Namenode:整个文件系统的管理节点。它维护着整个文件系统的文件目录树

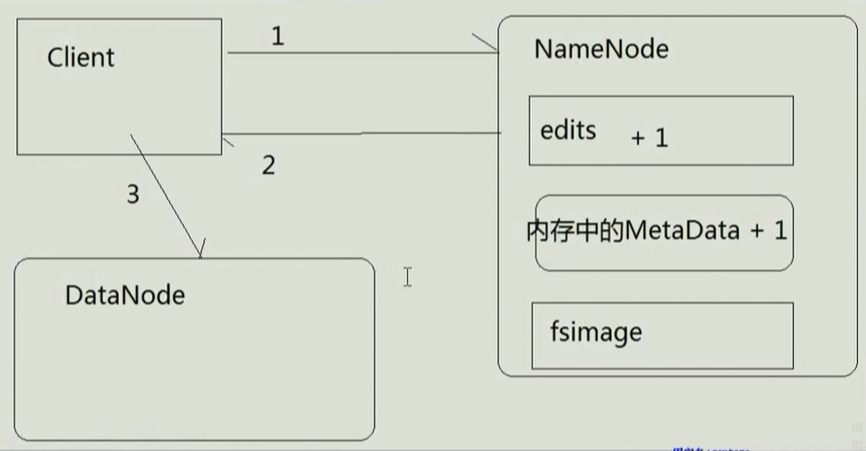




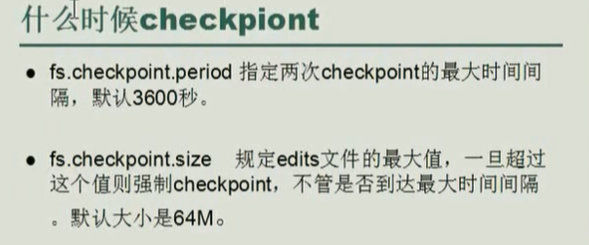








Checkpoint：检查点



实验：验证数据验证块的大小

$ ls –lh /home/tsing01/Downloads/

-rw-rw-r-- 1 tsing01 tsing01 411M Dec 8 08:47 hadoop-2.8.0.tar.gz

-rw-rw-r-- 1 tsing01 tsing01 174M Dec 8 08:21 jdk-8u111-linux-x64.tar.gz

-rw-rw-r-- 1 tsing01 tsing01 445 Dec 8 09:17 oo

$ Hadoop fs –put /home/tsing01/ jdk-8u111-linux-x64.tar.gz /

$ cd /usr/local/hadoop/tmp/dfs/data/current/BP-334601254-127.0.1.1-1544349873832/current/finalized/subdir0/subdir0

$ ls –lh

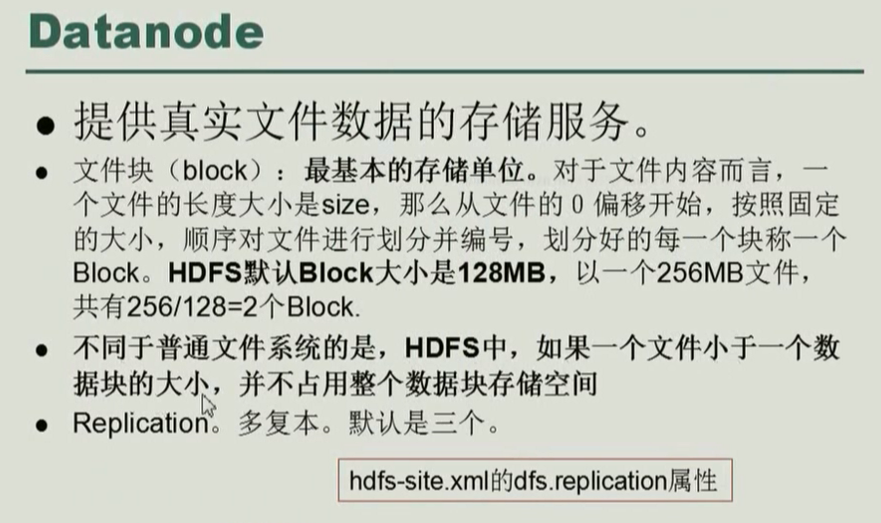
-rw-rw-r-- 1 hadoop hadoop 128M Dec 10 04:34 blk\_1073741827

-rw-rw-r-- 1 hadoop hadoop 1.1M Dec 10 04:34 blk\_1073741827\_1003.meta

-rw-rw-r-- 1 hadoop hadoop 46M Dec 10 04:34 blk\_1073741828

-rw-rw-r-- 1 hadoop hadoop 361K Dec 10 04:34 blk\_1073741828\_1004.meta

174M=128M+46M 最大为128M 小文件不是占用128M 而是原有大小



安装eclipse

版本：

Eclipse Java EE IDE for Web Developers.

Version: 2018-09 (4.9.0)

Build id: 20180917-1800

官网下载 双击安装 选择安装路径 选择工作空间

构建maven工程 http://www.cnblogs.com/Leo\_wl/p/4862820.html

编辑pom

<repositories><!—阿里代码库 -->

<repository>

<id>maven-ali</id>

<url>http://maven.aliyun.com/nexus/content/groups/public//</url>

<releases>

<enabled>true</enabled>

</releases>

<snapshots>

<enabled>true</enabled>

<updatePolicy>always</updatePolicy>

<checksumPolicy>fail</checksumPolicy>

</snapshots>

</repository>

</repositories>

<dependencies>

<dependency>

<groupId>org.apache.hadoop</groupId>

<artifactId>hadoop-common</artifactId>

<version>2.8.5</version>

</dependency>

<dependency>

<groupId>org.apache.hadoop</groupId>

<artifactId>hadoop-hdfs</artifactId>

<version>2.8.5</version>

</dependency>

<dependency>

<groupId>org.apache.hadoop</groupId>

<artifactId>hadoop-mapreduce-client-core</artifactId>

<version>2.8.5</version>

</dependency>

<dependency>

<groupId>org.apache.hadoop</groupId>

<artifactId>hadoop-mapreduce-client-jobclient</artifactId>

<version>2.8.5</version>

</dependency>

<dependency>

<groupId>org.apache.hadoop</groupId>

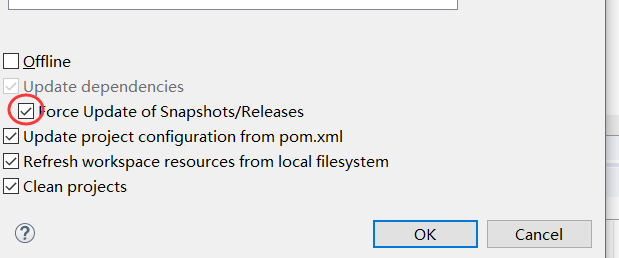
<artifactId>hadoop-mapreduce-client-common</artifactId>

<version>2.8.5</version>

</dependency>

</dependencies>

右键工程 – maven – update maven project



虚拟机ubuntu安装eclipse for linux

官网下载 ：

版本：

Eclipse Java EE IDE for Web Developers.

Version: 2018-09 (4.9.0)

Build id: 20180917-1800

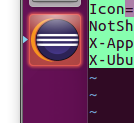
解压 tar –zxvf -C /usr/local/

进入目录 cd /usr/local/eclipse-installer

安装前需要确认虚拟机可以连接外网

安装 sudo ./eclipse-inst

将eclipse放到任务栏



# cd /usr/share/applications 进入编辑目录

# vim eclipse.desktop 新建eclipse.desktop配置文件并粘贴下面的代码

# gedit eclipse.desktop vim不支持粘贴复制的可以使用ubuntu自带的文档编辑器 gedit并粘贴下面的代码

[Desktop Entry]

Name=eclipse

Comment=javaee eclipse

GenericName=eclipse

Exec=/home/hadoop/eclipse/jee-2018-09/eclipse/eclipse #改为自己的安装路径

Terminal=false

Type=Application

Categories=Utility

Icon=/usr/share/unity/icons/eclipse.xmp #改为自己的图标路径

NotShowIn=GNOME-Flashback;

X-AppStream-Ignore=true

X-Ubuntu-Gettext-Domain=session-shortcuts

# source eclipse.desktop

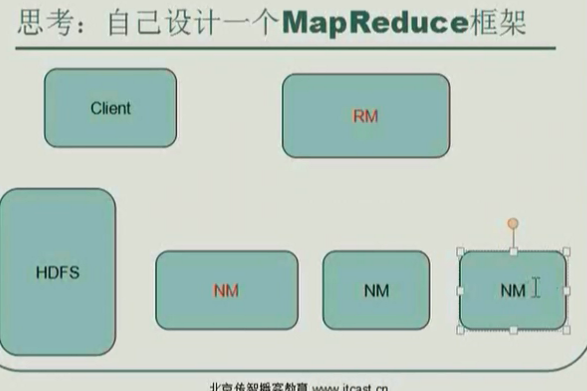
# 第三天 day3

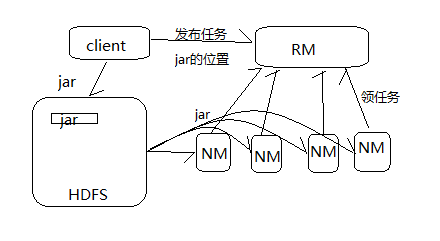
MapReduce

Map函数

Reduce函数

参数都是key：value形式

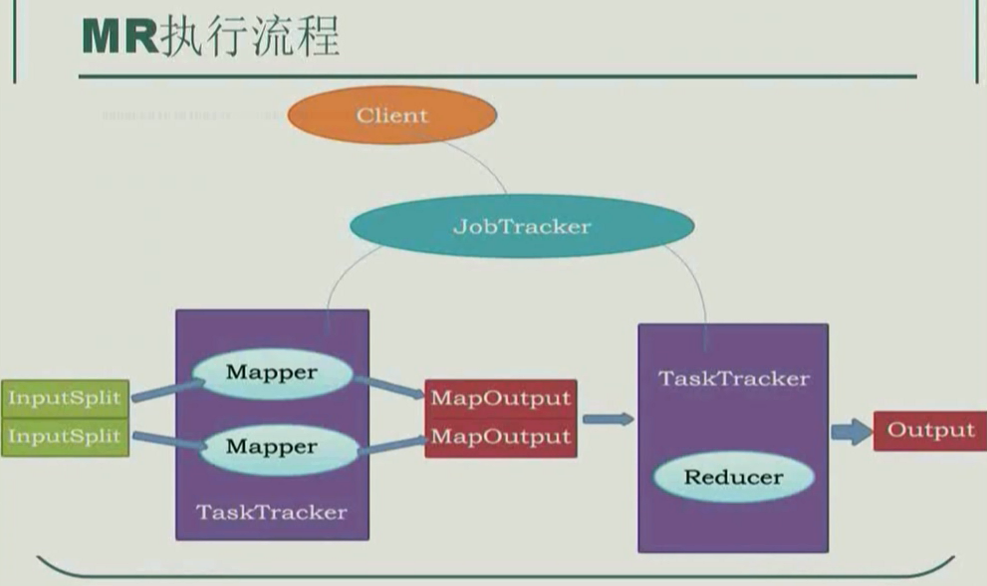


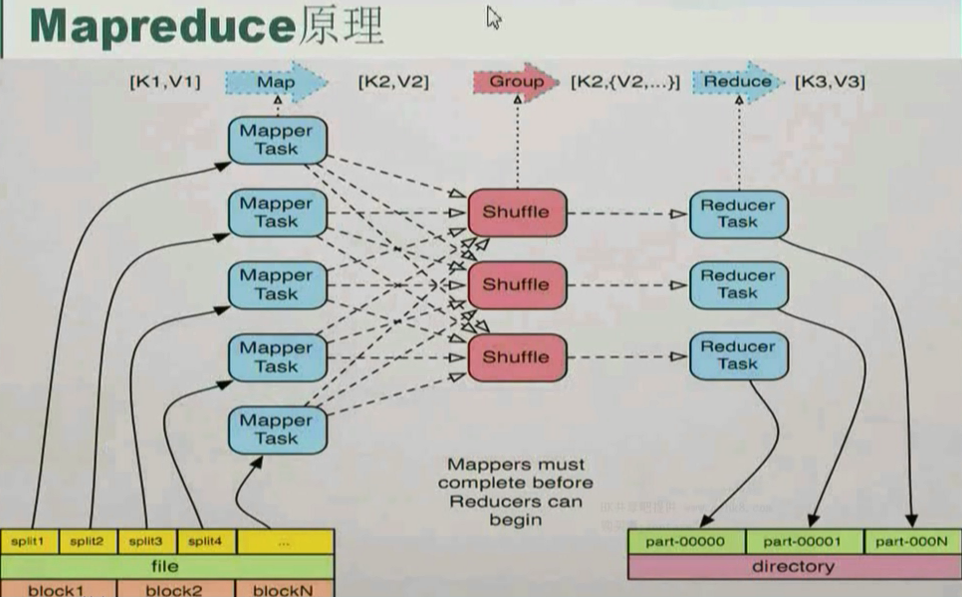


名称变化：

版本2：RM：resource manage 🡨版本1：job tracker

版本2：NM：node manage 🡨版本2：task tracker





Shuffle：

下载pdf <https://blog.csdn.net/yuan_xw/article/details/50003197>

# Hadoop集群搭建

教程，较为详细： <https://linux.cn/article-8695-1.html>

1、关闭防火墙：

查看当前防火墙状态

# ufw status

inactive状态是防火墙关闭状态

active是开启状态。

关闭防火墙

# ufw disable

2、配置 hostname

# hostnamectl set-hostname namenode ### 为 namenode 节点设置 hostname

# hostnamectl set-hostname datanode1 ### 为 datanode1 节点设置 hostname

# hostnamectl set-hostname datanode2 ### 为 datanode2 节点设置 hostname

Log存储位置

修改 hadoop log 存储位置

hadoop@namenode:/usr/local/hadoop/etc/hadoop$ echo "export HADOOP\_LOG\_DIR=~/log/hadoop" >>hadoop-env.sh

系统粘贴板不能使用：刚刚还是可以使用的

查找原因：

# vim

：reg

是否有 + 寄存器，如果没有 就重启mobaxterm 或者虚拟机

<https://blog.csdn.net/zhangxiao93/article/details/53677764>

无法建立连接：

The authenticity of host 'datanode1 (192.168.72.101)' can't be established.

ECDSA key fingerprint is SHA256:apttw6zRXnZFKX3O0L7ScHlXg2Bnb60oOZyqw/jucsY.

Are you sure you want to continue connecting (yes/no)? The authenticity of host 'datanode2 (192.168.72.102)' can't be established.

ECDSA key fingerprint is SHA256:apttw6zRXnZFKX3O0L7ScHlXg2Bnb60oOZyqw/jucsY.

Are you sure you want to continue connecting (yes/no)? localhost: datanode running as process 2086. Stop it first.

namenode: datanode running as process 2086. Stop it first.

datanode2: Host key verification failed.

datanode1: Host key verification failed.

原因推测：Permanently did not added 'namenode' (ECDSA) to the list of known hosts.即namenode没有添加到.ssh目录下的known\_host中。

解决：使用ssh各自连接一遍，将其添加到host\_knowns

hadoop@namenode:$ ssh namenode

hadoop@namenode:$ ssh datanode1

hadoop@namenode:$ ssh datanode2

hadoop@datanode1:$ ssh namenode

hadoop@datanode1:$ ssh datanode1

hadoop@datanode1:$ ssh datanode2

hadoop@datanode2:$ ssh namenode

hadoop@datanode2:$ ssh datanode1

hadoop@datanode2:$ ssh datanode2

向~/bashrc加入一句话

export PATH=:/usr/local/hadoop/bin:/usr/local/hadoop/sbin

导致系统无法使用ls vim 等基本命令

解决：

/usr/bin/vim ~/.bashrc 将上一句话改为

export PATH=$PATH:/usr/local/hadoop/bin:/usr/local/hadoop/sbin

重启电脑

Namenode没有启动：

原因：可能是tmp导致的文件不一致

解决：改变存放dfs的位置

Vim core-site.xml

<property>

<name>hadoop.tmp.dir</name>

<value>/usr/local/hadoop/hadoop\_tmp</value>

</property>

Mobaxterm 远程ubuntu无法使用ls ，但是ubuntu自己可以使用ls？？？

晚上回家就好了，也没有关机。神奇

linux如何查看端口占用

1、lsof -i:端口号

2、netstat -tunlp|grep 端口号

集群总览：

| **NodeName** | **Web** | **Address** |
| --- | --- | --- |
| namenote | HDFS | <http://192.168.72.100:50070> |
|  | RM | <http://192.168.72.100:8088> |
|  | NM | <http://192.168.72.100:8042> |
| datanode1 | HDFS | <http://192.168.72.101:50075> |
|  | NM | http://192.168. 72.101:8042 |
| datanode2 | HDFS | <http://192.168.72.102:50075> |
|  | NM | http://192.168. 72.102:8042 |

# RPC 未完成

根据该教程编程https://blog.csdn.net/u012453843/article/details/52496514

验证端口9000的可用性 ?？？

# Telnet namenode 9000

https://blog.csdn.net/msq7487223/article/details/52366148

RPC介绍 <http://www.cnblogs.com/flyoung2008/archive/2011/11/12/2246282.html>

杀死占用端口port的进程：

1、找到使用该端口的进程的pid：最后一栏的数字

>Netstat –ano | findstr port

2、taskkill命令杀死进程

>taskkill taskkill /f /t /im pid

\*\*未结束\*\*

MapReduce：需要导入包 。。。。。

# Hadoop序列化机制

---------------------

作者：在京奋斗者

来源：CSDN

原文：https://blog.csdn.net/u012453843/article/details/52600313

版权声明：本文为博主原创文章，转载请附上博文链接！

为什么不使用java的序列化机制？

用我们的java序列化机制序列化Monkey这个类时，还要保存它与父类Animal的继承结构。反序列化时当然这些也是需要保留的。我们Hadoop要序列化某个类是根本不需要保留它们的继承结构的，只需要能够实现持久化，能够实现快速传递就可以了。这样的话，Java的序列化就会非常冗余，因此Hadoop没有采用Java的序列化机制，而是自己实现了一套序列化机制。

Hadoop序列化的特点如下：

1. 紧凑，高效使用存储空间

2. 快速，读写数据的额外开销小

3. 可扩展性，可透明地读取老格式的数据

4. 互操作，支持多语言的交互

WordCount.class？？？？？？？？？？？

# 学习Hadoop第十一课（MapReduce的实例---WordCount）

出现的错误类型与示例一样，但是结果不同

我的

hello 1

jerry 1

kitty 1

tom 1

word 1

示例的

hello 5

jerry 1

kitty 1

tom 2

word 1

**datacount**

数据格式

访问时间 手机号或上网卡 网络运营商的MAC地址 上网机器的IP 访问网站 网址类型 上行包 下行包 上行流量 下行流量 状态

1363157985066 13726230503 00-FD-07-A4-72-B8:CMCC 120.196.100.82 www.taobao.com 购物网站 24 27 2481 24681 200

1363157995052 13826544101 5C-0E-8B-C7-F1-E0:CMCC 120.197.40.4 4 0 264 0 200

1363157991076 13926435656 20-10-7A-28-CC-0A:CMCC 120.196.100.99 2 4 132 1512 200

把首行去掉

上面的数据对应的字段名称、意义及列索引号如下图所示。



写完MapReduce 打jar包 上传到ubuntu 运行

hadoop jar dc.jar com.myhadoop.dc.DataCount /liuliang /dc4

出错

java.lang.Exception: java.lang.ArrayIndexOutOfBoundsException: 1

at org.apache.hadoop.mapred.LocalJobRunner$Job.runTasks(LocalJobRunner.java:489)

at org.apache.hadoop.mapred.LocalJobRunner$Job.run(LocalJobRunner.java:549)

Caused by: java.lang.ArrayIndexOutOfBoundsException: 1

at com.myhadoop.dc.DCMapper.map(DCMapper.java:23)

at com.myhadoop.dc.DCMapper.map(DCMapper.java:1)

at org.apache.hadoop.mapreduce.Mapper.run(Mapper.java:146)

打完jar包再调试 不对再打jar包 调试 太麻烦。

卡在调试hadoop上1天了

思路1、 win10 eclipse安装hadoop插件实现本地调试

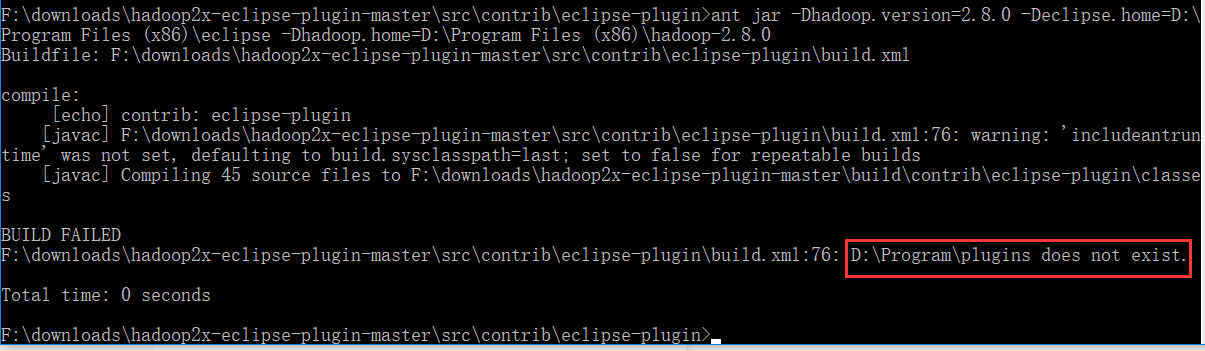
插件安装

问题1 下载别人hadoop插件 jar 是他们是自己编译的

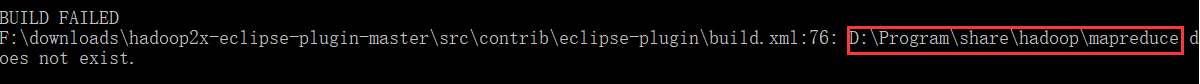
问题2 如何自己编译hadoop插件 jar。

<https://blog.csdn.net/jy02268879/article/details/80273683>

用win10 eclipse安装遇到的问题：



手动新建文件夹

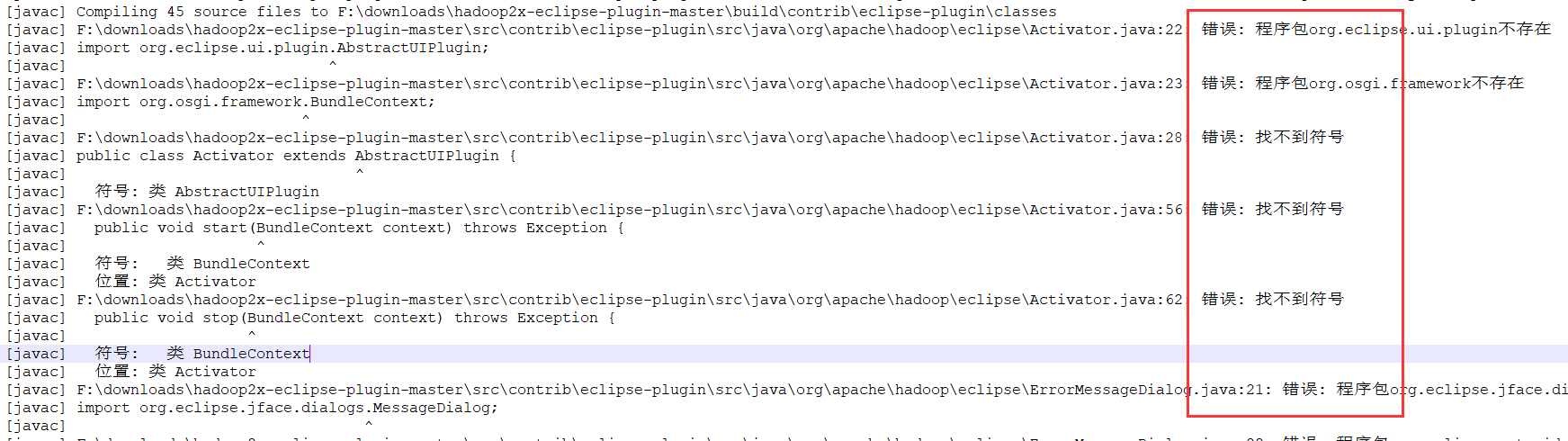


我的编译命令

ant jar -Dhadoop.version=2.8.0 -Declipse.home=D:\Program Files (x86)\eclipse -Dhadoop.home=D:\Program Files (x86)\hadoop-2.8.0

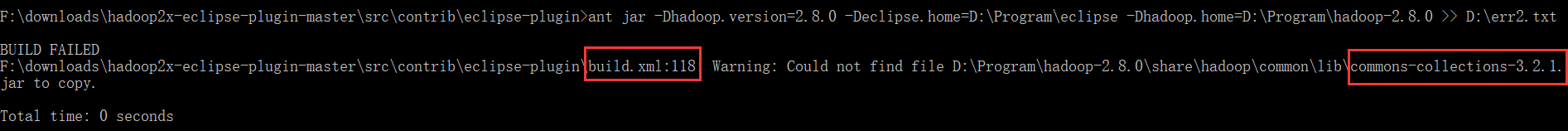
是空格导致的错误

将hadoop移动到D:/Program 重新编译



是空格导致的错误：hadoop和eclipse的路径中不能有空格，否则ant无法编译

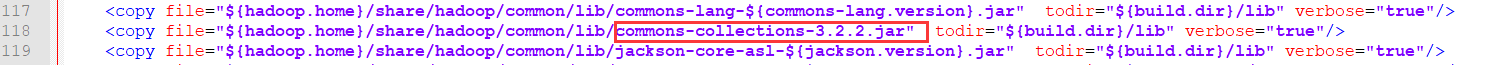
将eclipse移动到D:/Program 重新编译



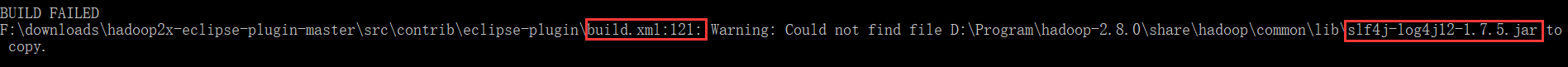
根据上图中第二个红框前面的位置找jar包

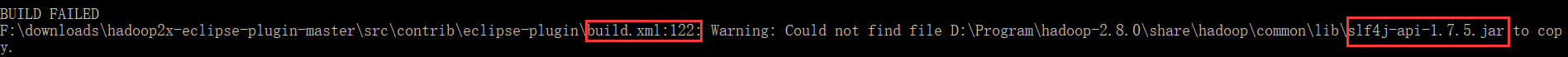


已经从3.2.1升级到3.2.2，手动修改build.xml 118行就好



又有两个个jar包找不到





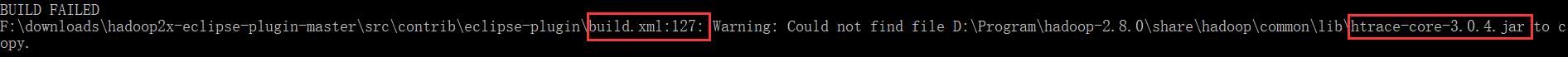
根据上图中第二个红框前面的位置找jar包



修改build.xml 121 122行



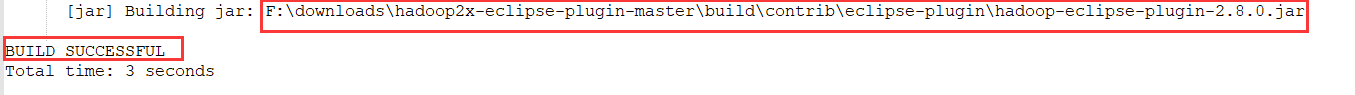
还有jar找不到



方法同上：







成功编译hadoop插件

将hadoop-eclipse-plugin-2.8.0.jar放到eclipse/plgins

windows》open perspective还是找不到MapReduce

Ubuntu + Hadoop编译插件

<https://www.cnblogs.com/yts1dx/p/4020263.html>

# apt install ant #安装ant

# ant –version #验证是否成功

Apache Ant(TM) version 1.9.6 compiled on July 20 2018

手动安装

到Apache官网下载最新版本的ant：<http://ant.apache.org/bindownload.cgi>

下面的步骤注意自己的版本号

  # tar -xf apache-ant-1.8.2-bin.tar.gz #(可能会要求输入密码)

解压下载下来的.tar.gz文件：

# mv apache-ant-1.8.2 /opt/ #将解压出来的文件移动到/opt/下

# vim /etc/profile #配置环境变量，在原来基础上添加以下蓝体字：

     export ANT\_HOME=/opt/apache-ant-1.8.2  
     export JAVA\_HOME=/usr/lib/jvm/java-6-openjdk  
     export PATH=$JAVA\_HOME/bin:$PATH:$ANT\_HOME/bin  
     export CLASSPATH=.:$JAVA\_HOME/lib/dt.jar:$JAVA\_HOME/lib/tools.jar

  # ant –version #验证是否安装成功：

Apache Ant(TM) version 1.8.2 compiled on December 20 2010  
    如此字样，则表示安装成功！

具体的之用方法，使用 ant --help 即可一一学到。

编译源码

将源码文件hadoop2x-eclipse-plugin-master.zip和ant传送到/root

打开terminal

# unzip hadoop2x-eclipse-plugin-master.zip #zip解压到当前路径

# ~/hadoop2x-eclipse-plugin-master/src/contrib #进入build.xml所在目录，否则直接编译报错 build.xml does not exist！

# ant jar -Dversion=2.8.0 -Declipse.home=/home/hadoop/eclipse/jee-2018-09/eclipse/eclipse -Dhadoop.home=/usr/local/hadoop

报错：

BUILD FAILED

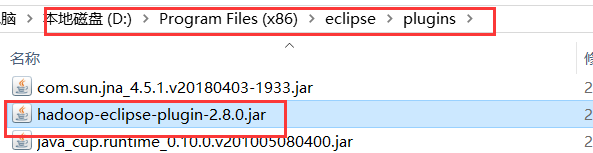
Target "jar" does not exist in the project "hadoopcontrib".

步骤1、下载别人编译好的jar

https://github.com/larntin/hadoop2x-eclipse-plugin/tree/v2.8.0/release

hadoop-eclipse-plugin-2.8.0.jar

步骤2、将hadoop-eclipse-plugin-2.8.0.ja放入D:\Program Files (x86)\eclipse\plugins（D:\Program Files (x86)\eclipse\是eclipse安装目录）



步骤3、重启eclipse

无法在Open PersPective中找到Map/Reduce插件 。。。。。。

解压hadooptar到D:/program files/

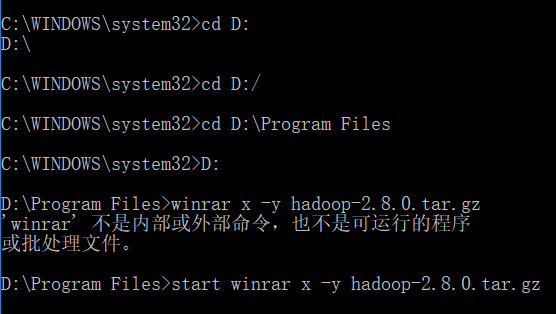
普通的右键》解压到\*\*\*不行，提示无法创建符号链接 E:\hadoop\hadoop-3.0.3\hadoop-3.0.3\lib\native\libhadoop.so您可能需要以管理器身份运行 WinRAR

https://blog.csdn.net/u010674004/article/details/82117430

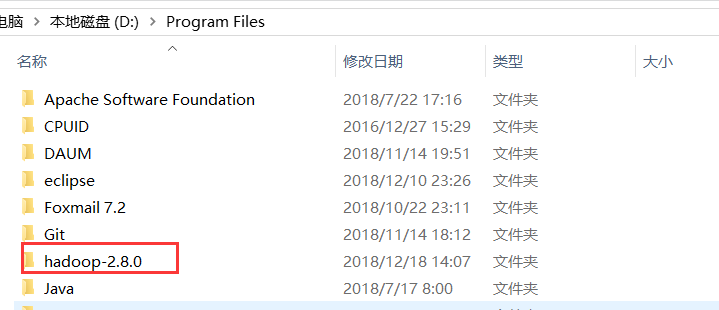
单击windows键》输入cmd



右键单击》以管理员身份运行；



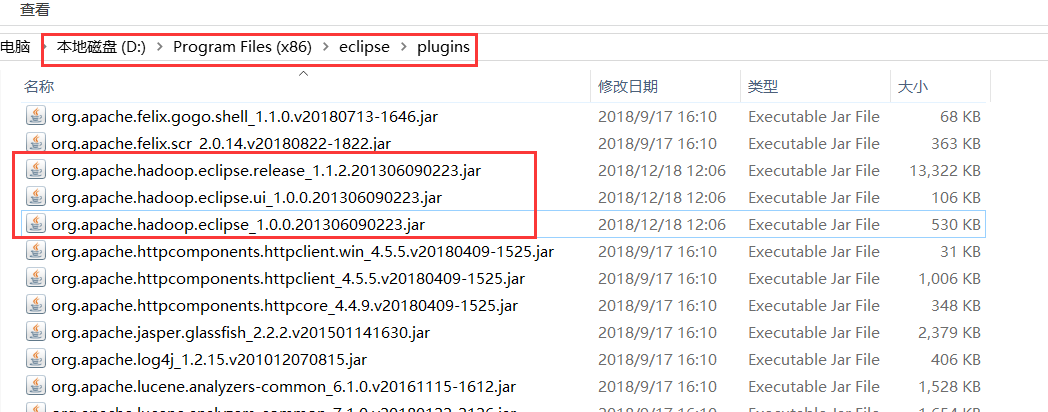
解压成功



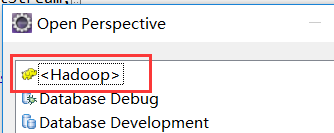
下载了<http://people.apache.org/~srimanth/hadoop-eclipse/update-site/plugins/>上的三个jar



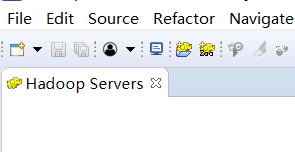
放入D:\Program Files (x86)\eclipse\plugins（D:\Program Files (x86)\eclipse\是eclipse安装目录）

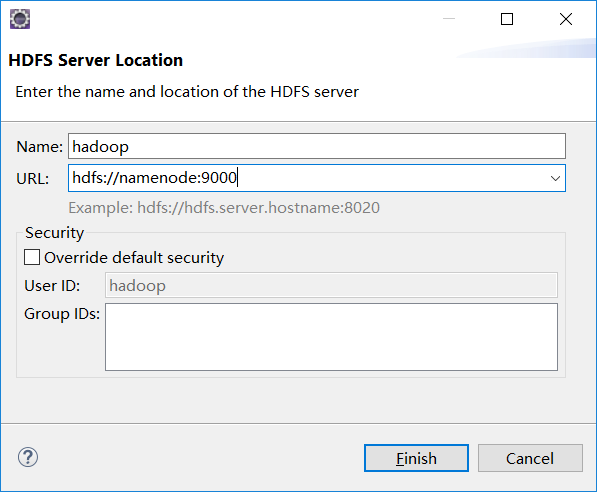


然后windows 》perspective》open perspective》others 多出来一个hadoop



打开后多出hadoop servers





出现错误：Server IPC version 9 cannot communicate with client version 4

原因：服务器是hadoop版本与 而本地eclipse开发环境的项目依赖的是hadoop-client版本不一致

解决：pom中配置hadoop-client

<dependency>

<groupId>org.apache.hadoop</groupId>

<artifactId>hadoop-client</artifactId>

<version>${hadoop.version}</version>

</dependency>

出现错误：

An internal error occurred during: "Creating HDFS project [hadoop]".

java.lang.NullPointerException

思路2、 ubuntu安装eclipse 调试

遇到 自动停止的bug

没有日志 无法查看详情 预测java问题。

1、将java hedoop的PATH放入/etc/profile(全部用户可用) 排除用户问题

2、

2天没有完成bug调试任务